Towards Evaluation of Personalized and Collaborative Information Retrieval

Debasis Ganguly, Johannes Leveling, Wei Li, and Gareth J.F. Jones

CNGL, School of Computing
Dublin City University
Dublin 9, Ireland
{dganguly, jleveling, wli, gjones}@computing.dcu.ie

Abstract. We propose to extend standard information retrieval (IR) ad-hoc test collection design to facilitate research on personalized and collaborative IR by gathering additional meta-information during the topic (query) development process. We propose a controlled query generation process with activity logging for each topic developer. The standard ad-hoc collection will thus be accompanied by a new set of thematically related topics and the associated log information, and has the potential to simulate a real-world search scenario to encourage retrieval systems to mine user information from the logs to improve IR effectiveness. The proposed methodology described in this paper will be applied in a pilot task which is scheduled to run in the FIRE 2011 evaluation campaign. The task aims at investigating the research question of whether personalized and collaborative IR retrieval experiments and evaluation can be pursued by enriching a standard ad-hoc collection with such meta-information.

1 Introduction

One major challenge in Information Retrieval (IR) is the potential to adapt a retrieval model for personalized IR. Different users may enter the same query string into a search system, but their information needs can be vastly different. The notion of relevance depends upon factors such as the domain knowledge of the searcher, information gained from reading previous documents in the past, and general search behavior of a searcher, e.g. how many documents he normally reads before reformulating his search [1].

In a typical laboratory evaluation scenario of ad-hoc IR, participants are given a document collection and a set of queries (topics). The task of the participating systems is then to retrieve documents which satisfy the information need expressed in each query. Such a traditional evaluation framework does not provide enough information to facilitate personalized IR. This information includes: a) closely related topics formulated by different people with different assessments reflecting a differing notion of relevance, and b) meta-information such as the documents viewed by the users.

The process of TREC-style topic development is artificial and does not resemble iterative query reformulation in real search activities where typically a user of the search system enters an initial query, reads a few top ranked retrieved documents and reformulates the initial query. The final query, based on the content read thus far, retrieves one or more relevant items which satisfy his information need up to this point. Our

main hypothesis is that this iterative process of topic development is more similar to the real-world search than a search based on a single topic.

To our knowledge, little or no research has addressed gathering and providing metadata for the query development process under the framework of an ad-hoc retrieval dataset.[1] There is no freely available dataset for research purposes containing search history logs on a closed document collection. Our work attempts to provide a common evaluation framework to test various personalized IR systems.

The rest of the paper is organized as follows: Section 2 surveys work on user modelling and personalized IR, Section 3 presents our approach to generating user logs in a controlled environment, Section 4 outlines the planned task to be undertaken within the FIRE 2011, and Section 5 concludes the paper with a brief summary and outlook.

2 Related Work

The research question we want to explore is whether IR systems can present more relevant documents to individual searchers (hence addressing personalization) by exploiting his own browsing information and that of users with similar search interests.

Recent works on the study of user search patterns include that of Kellar et. al. [5]. They report that users spend most of their time, view most pages, and extensively use the browser functions for information gathering tasks, thus establishing the need for extensive user studies of information gathering tasks. Kelly and Belkin [6] report that there is no direct relationship between the display time of a document and its usefulness, and that display times differ significantly according to a specific task and according to individual users. White and Kelly [10] show that tailoring document display time thresholds for implicit relevance feedback based on task grouping is beneficial for personalization. Liu and Belkin [8] design a method for decomposing tasks into sets of (in)dependant subtasks and show that documents viewed in previous searches can help searchers to find useful documents on a related topic. This is attributed to the fact that users gain knowledge across stages regarding the usefulness of documents.

The related work on user studies shows that i) useful user log information can be gathered by designing information gathering tasks; ii) document display times can be used for implicit relevance feedback to benefit retrieval models; and iii) information regarding the previous visited documents can probably be utilized by retrieval systems to model the changing notion of relevance. It particularly encourages us to generate a log of the entire topic creation process to make all information about the search process available to the retrieval systems, which has the potential to help tuning IR systems to user-specific needs.

The LogCLEF[2] log analysis initiative provides log data from different providers [2], but these datasets lack information about query variants on the same topic by different users and do not include relevance assessments.

TREC 2010 introduced a new track called the Sessions Track[3], where the motivation is to form and evaluate a session of related queries [4]. This track involves modifying a

[1] Metadata includes all information from the search history for all query formulations.

[2] http://www.promise-noe.eu/mining-user-preference/logclef-2011/

[3] http://ir.cis.udel.edu/sessions/

starting query into a more general query, a more specific query, or one addressing another facet of the information need. Our proposed track is different because firstly, we do not manually form query variants, but expect the participants to contribute in generating search data and provide them with search logs from other participating and volunteering topic developers. Secondly, our track is not primarily concerned with query sessions, but with categorizing users based on their interests and with exploring whether individual searchers can profit from information about similar searches or users.

NTCIR-9[4] is organizing the Intent task, where topics are formed automatically by random sampling from Chinese web search query logs. A difference between our proposed task and the NTCIR Intent task is that the latter deals with web search and uses a bottom-up approach (starting from existing query logs), whereas we try to address elements of personalization with a top-down approach, aiming to create interaction logs.

In summary, there are two important differences compared to previous research: 1) The topic development and relevance assessment will be performed by the same person. 2) The same (static) corpus is utilized for search and logging the topic development process, because experiments which are based on web search logs are typically not reproducible due to the dynamic nature of web documents.

3 Proposed Methodology

Our proposed methodology is aimed towards achieveing the following objectives: i) generation of logs; ii) analysis of the logs; and iii) report and analyze the effect of this meta-information on retrieval effectiveness. To promote our approach to automatic "closed-box" personalized and collaborative IR experiments and encourage researchers to use and contribute to this method, we proposed our methodology of extending a standard test collection with user-log meta-information to the Forum of Information Retrieval and Evaluation[5] (FIRE). This proposal has been accepted and will run as a pilot task in FIRE 2011.[6] We plan to employ the following methodology for compiling a log dataset for the FIRE ad-hoc English collection. A web interface will be developed and hosted, which will be used at all stages of the pilot task. A topic developer (a volunteer or a participant interested to contribute to the topic development) will log into the system with a registered user ID. The system logs all user actions. The topic creators then go through a search phase (selecting the search category, submitting queries and viewing result documents) and a topic formulation and evaluation phase (summarizing the found information, formulating the final topic, and assessing relevance for documents). The main task in the last phase is to compile information from different documents covering different aspects of the category. The topic development procedure is illustrated in Fig. 1:

1. Category selection (topic developer). The system presents a list of broad search *categories* from which the topic developer has to select one. Deriving the topics from a pre-defined list of categories is intended to ensure development of related topics with overlapping information needs for different users.

[4] http://www.thuir.org/intent/ntcir9/
[5] http://www.isical.ac.in/~clia/
[6] http://www.cngl.ie/Fire-PIR/

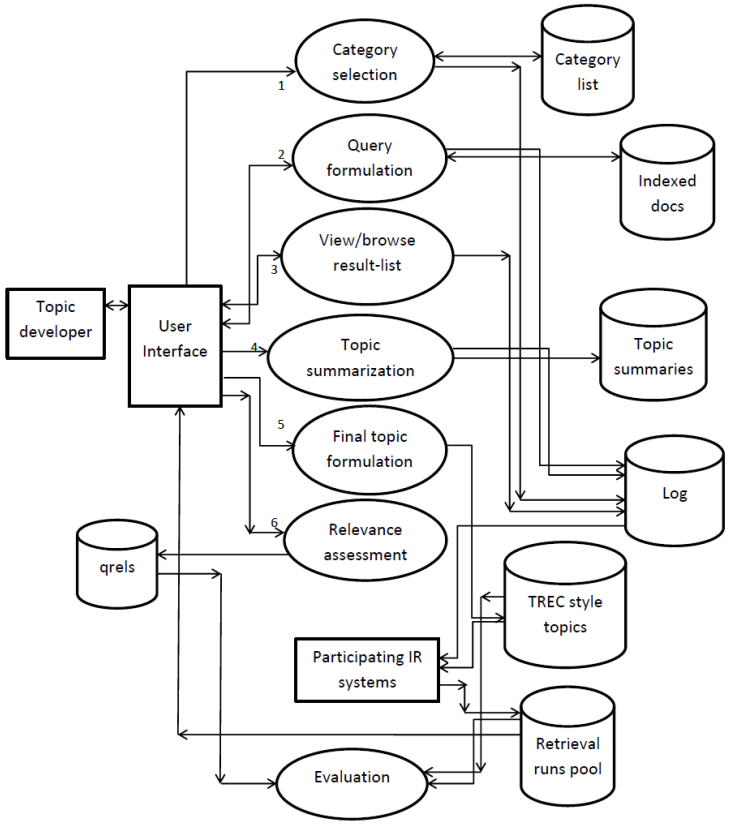

Fig. 1: Data flow diagram of the topic development phase.

The search categories will be selected in accordance to the TREC guidelines of topic development, which involves performing trial retrievals against the document set and choosing topics for which the result set is not too small or too large [3]. To ensure a roughly uniform distribution of queries across the search categories, the web interface will limit the available categories according to their selection frequency.

2. Query formulation and retrieval. After selecting a category, the user will iterate through query formulations, retrieving different documents at each iteration.

3. View/browse result documents. The user will read documents retrieved in the previous step from the FIRE ad-hoc document collection by an IR system and bookmarks a few before he feels that he has gained sufficient knowledge about the topic.

4. Topic summarization. The system presents a form where the topic developer has to enter a report/summary on the current topic. We view this summary as a means to ensure that the topic developer has gained knowledge about the search category and that the final topic is based on information from documents viewed by the topic developer.

5. Final topic formulation. As a next step the system asks them to form a TREC-formatted topic based on the knowledge gained thus far. This query aims at one user-specific, "personalizable" aspect of the initial search category, i.e. one particular aspect of the category that the topic developer is especially interested in. The topic developers have to fill in the *title*, *description* and *narrative* fields for the query, describing the information need by a phrase, a full sentence, and a description of which documents are relevant and which are not. These TREC-style topics serve as input for the IR runs.

6. Relevance assessment. Relevance assessments are based on the pool of submissions. The developer of a topic will be assigned the responsibility to mark the relevant documents according to the relevance criteria expressed in the narrative field of the topic provided by him. An interesting observation to be made here is to see whether there exists a *personal* notion of relevance, i.e. how often a document is relevant for two different topics belonging to the same category. Another observation to be made is to see how many of the documents bookmarked or viewed for a long time by topic developers for a category are actually relevant for the topics in that category. A higher number would justify the use of logging information as a means for supporting relevance assessment with information gathered during the topic development process.

For conclusive retrieval experiments, we expect at least three queries to be submitted to the system from every topic developer. We would require at least 10 topic developers using the system, and expect to have 30 submitted topics in TREC format with valid *title*, *description* and *narrative* fields.

3.1 An Example Scenario

Let us assume a topic developer selects the example topic "Impact of strikes" from the pre-defined list of search categories. He then enters a series of queries in the system (e.g. "Bengal strikes", "Violent protests strike", "Union strikes"), views the documents, bookmarks some of them and starts gaining knowledge about the category given to him. Figure 2 illustrates this topic development process. For the chosen initial search category, the user issues a query Q_1 and gets a ranked list of returned documents $\{D_1^1, \ldots D_m^1\}$. A subset of relevant documents (viewing or bookmarking a document might be an implicit indicator of relevance) is used to reformulate Q_1 into Q_2. The released meta-information corpus would thus contain each intermediate query Q_i, the set of top documents returned for Q_i namely $\{D_1^i, \ldots D_m^i\}$ and the actions of the user.

After going through a few iterations, the user then fills up the topic summary and submits his topic titled "What social effects do strikes have on the public life of Bengal?" with an appropriate description and narrative fields. Later on he also has to assess documents for relevance for this topic.

Consider another topic developer who selects the same topic. He also browses documents through the system and eventually ends up with a topic titled "Impact of strikes for employees in the Information Technology sector".

If the first query belongs to the *training* set and the second to the *test* set, the challenge for the participants is firstly to identify (using the query logs or other external information) that both these queries belong to the same search category which suggests that these two topic developers belong to the interest group. Information about one user might benefit satisfying the information need(s) of the other. The next challenge for

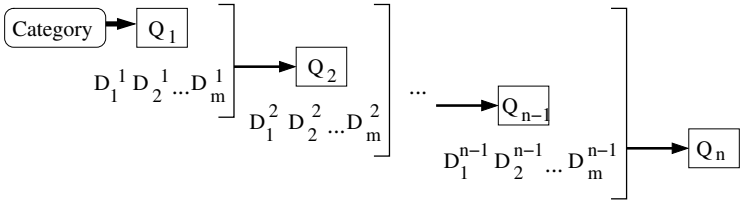

Fig. 2: Structure of the meta-information.

the participants is to develop ideas of how to increase retrieval effectiveness for both searchers by exploiting the browsing history of both users.

3.2 Task Description

Training phase. One third of the submitted topics will be released as *training* topics which contain the search category selected while developing the topic. This will give the participants an opportunity to train their classification systems for predicting the search category.
Testing phase. The remaining two thirds comprise the *test* topics where the search category for each topic will not be disclosed. The objective here is to see how accurately IR systems can predict the broad category and use this information to form user groups. We hypothesize that if different individual users select the same category, they have similar interests and although they might end up forming different queries, these queries would still serve some overlapping information need.
Data details. Participants of this track will be provided with the following data:

A) Document collection - English FIRE ad-hoc collection. Depending on the number of interested participants, the task could also be offered for Indian languages such as Bengali or Hindi used within other FIRE tracks.
B) Browsing logs in CSV format containing user ID, timestamp, and details of the action performed by the user, which is one of:
   1. Category selection,
   2. Query formulation and reading results,
   3. View/browse result documents,
   4. Topic summarization,
   5. Final topic formulation,
   6. Relevance assessment.
C) Queries in TREC format formed by the process outlined in Section 3. All queries (*training* and *test*) will have one additional field - *user_id* which is the registered user ID of the topic developer. The *training* topics will have another additional field - *category* for the training topics.

In addition to the above, relevance judgments in TREC format, assessed by the topic developers will be made a part of the collection after the evaluation results are released.

4 Proposed tasks

We propose two tasks to be run under this track with an aim to answer the following research questions: a) the degree of accuracy to which systems can predict topic categories based on the query logs; and b) the degree of improvement gained in retrieval effectiveness by utilization of predicted user intents.

4.1 Category Finding Task

Participants are required to predict the top-level search category for each *test* topic utilizing the browsing history of the current user (i.e. the viewed documents). It is expected that if two topic developers have selected the same topic, there is a similarity in their general interests. This task will address the question of how effectively systems can classify an unknown query into one of the search categories.

This task is somewhat similar to the filtering task of TREC where a given document has to be classified into one of the existing categories of documents [7]. Rose and Levinson [9] categorized web queries into navigational, informational, and transactional types and advocate the necessity of search engines to predict the user intent for addressing personalization. Our work is an attempt to explore whether a finer level of intent prediction (an intent in our scenario corresponds to the top level search category) actually benefits the retrieval systems for ad-hoc IR.

The basic objective of this task is to measure the degree of accuracy to which systems can correctly recognize the category of the topics. The problem can be mapped to a standard multi-class classification problem where the topic categories are the class labels and each test topic needs to be assigned to one of these classes. Thus we plan to use standard classification metrics such as precision, recall and ROC curves for evaluation.

4.2 IR Task

After the category finding task, we will release the true categories of the topics. This will allow research groups who are not participating in the topic category finding task to participate in the personalized IR task only. The IR task involves tuning retrieval systems to address individual user-specific information needs using the predicted search categories of the individual users. The participants will be asked to submit at least two out of the following three retrieval runs:

- `BL` (Baseline): a baseline ad-hoc run without using the logs and user group information. This is a mandatory retrieval run for all participants.
- `PPIR` (Predicted Personalized IR): a run involving user group models constructed by the predicted topic categories.
- `TPIR` (True Personalized IR): a run utilizing true topic categories for grouping users.

Participants are encouraged to derive user models from the additional user information (search history etc.) provided by us. We also plan to provide a baseline implementation for each subtask.

This task would use standard IR evaluation metrics such as MAP and P@10. The objective here is to see whether the associated metadata can be utilized to benefit personalized IR. The degree of improvement by automated analysis of the provided metadata is given by the relative difference of MAP(PPIR) with respect to MAP(BL). Through the retrieval performance of TPIR we will know the gold-standard that can be achieved in retrieval effectiveness when the true topic categories are given. Our expected observation is MAP(TPIR) $>$ MAP(PPIR) and MAP(PPIR) $>$ MAP(BL).

5 Conclusions and Outlook

This paper reports our plan for building a test data corpus for personalized IR, the tasks planned for the participants and the evaluation methodologies for a new pilot track to be run at FIRE 2011. All FIRE ad-hoc track participants can participate in this track. Participants from related tracks such as the Sessions Track at TREC, LogCLEF at CLEF, and the intent finding track at NTCIR could also be interested to participate in this track. The task is planned for English, but as the methodology is language-independent, it can be applied for Indian languages also used at FIRE (e.g. Bengali or Hindi) if enough interest can be raised from FIRE participants.

The proposed methodology can act as the first stepping stone towards evaluation of different retrieval systems under the same test bed of user generated logs. The log generation process has been designed to address aspects of personalization by capturing individual information needs for a broad search category. The history of the documents viewed prior to developing the final topic makes the topic development process transparent to a retrieval system.

Acknowledgments

This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie).

References

1. M. J. Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13(5):407–424, 1989.
2. G. M. Di Nunzio, J. Leveling, and T. Mandl. Multilingual log analysis: LogCLEF. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Murdoch, editors, *ECIR 2011*, volume 6611 of *LNCS*, pages 675–678. Springer, 2011.
3. D. Harman. Overview of the third text retrieval conference (TREC-3). In *TREC*, 1994.
4. E. Kanoulas, P. Clough, B. Carterette, and M. Sanderson. Session track at TREC 2010. In *SIMINT workshop SIGIR '10*. ACM, 2010.
5. M. Kellar, C. R. Watters, and M. A. Shepherd. A field study characterizing web-based information-seeking tasks. *JASIST*, 58(7):999–1018, 2007.
6. D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *SIGIR '04*, pages 377–384. ACM, 2004.
7. D. D. Lewis. The TREC-4 filtering track. In *The Fifth Text REtrieval Conference (TREC-5*, pages 75–96, 1997.

8. J. Liu and N. J. Belkin. Personalizing information retrieval for multi-session tasks: the roles of task stage and task type. In *SIGIR'10*, pages 26–33. ACM, 2010.
9. D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04*, pages 13–19, New York, NY, USA, 2004. ACM.
10. R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *CIKM 2006*, pages 297–306. ACM, 2006.